

一种改进的模糊人工免疫网络数据分类方法

苟世宁, 杜海峰, 栗茂林, 庄健

(西安交通大学现代设计和转子轴承系统教育部重点实验室, 710049, 西安)

摘要: 针对现有人工免疫网络算法对先验知识应用不足的问题, 提出一种基于模糊人工免疫网络的有监督学习数据分类方法. 首先采用模糊C均值聚类算法为免疫网络提供疫苗(初始种群), 将此疫苗作为免疫网络的初始抗体群, 种群再经过克隆选择、网络压缩、免疫成熟、记忆等算子的不断扩展和压缩, 形成一个由浓缩后的训练数据构成的抗体网络, 最终基于该抗体网络采用“邻近原则”构造分类器. 由于各算子的协调作用, 该方法能够在高浓缩率的情况下更好地代替样本空间. UCI (University of California, Irvine) 数据集的仿真实验证明, 与 aiNet 方法相比, 该方法在分类准确率和数据浓缩率上分别高出 7.26% 和 11.16%, 而且更稳定、可靠.

关键词: 人工免疫网络; 数据分类; 模糊

中图分类号: TP183 **文献标识码:** A **文章编号:** 0253-987X(2007)05-0585-04

Data Classification Based on a Modified Fuzzy Artificial Immune Network

Gou Shining, Du Haifeng, Li Maolin, Zhuang Jian

(Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System,
Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Based on fuzzy theory and artificial immune network, a new data classification method, named fuzzy artificial immune network classification (FAINC), is put forward. In order to improve the convergence rate, fuzzy C mean clustering algorithm is employed to provide vaccines (initial population) for artificial immune network. The operators such as clonal selection, network compression, immune maturation and immune memory are explored. According to expanding and compressing the population of the network, a stable antibody network can be acquired. Namely, the antibody network represents the concentrated training data to construct the classifier. The simulation experiments of University of California, Irvine (UCI) data sets validate the higher classification accuracy, data enrichment, stability and reliability of FAINC than aiNet.

Keywords: artificial immune network; data classification; fuzzy

分类是数据挖掘技术的重点内容之一, 在机器学习和模式识别等方面都有广泛的应用. 数据分类就是根据数据集的特点构造一个分类器, 利用分类器对未知类别的样本赋予类别的一种技术, 文献[1]认为常见的一些分类方法都或多或少地存在一些缺陷, 如 C4.5 算法效率低, 不适用于大数据集, 基于 Bayes 的分类算法都是以独立性假设为前提的, 而现实中的数据很难满足这个假设等. 随着人工智能

技术的不断发展, 遗传算法^[2]、神经网络^[3]和人工免疫系统^[4]等新方法被引入数据分类, 这些方法通过对数据集的不断搜索或学习力图找到全局最优的分类规则, 它们的效率比较高, 并且可以用于规模较大的数据集.

人工免疫系统是 20 世纪 90 年代产生的一种新的人工智能技术, 它模仿生物(主要是人)的免疫系统, 具有学习、记忆、分布式处理等功能, 为信息处理

提供了新的方法^[5],已经应用于控制^[6]、聚类^[7]和计算机安全^[8]等领域.人工免疫网络具有很好的容噪能力、泛化能力、记忆能力以及通过竞争实现并行分布处理的能力^[9],所以它能够对任意形状的数据进行浓缩,同时不受噪声信息的影响,为构造基于数据浓缩的分类器提供了新思路,是人工免疫系统研究的一个重要分支.目前,两个比较有影响的免疫网络模型是 Timmis 的资源受限人工免疫系统(Resource Limited Artificial Immune System, RLAIS)^[10]和 Castro 的 aiNet^[7].文献[11]在结合 RLAIS 和 aiNet 的基础上提出了一种用于聚类分析的进化模糊人工免疫网络(EFAIN),与 aiNet 相比,EFAIN 具有更高的数据浓缩率,但是和 RLAIS 及 aiNet 一样,EFAIN 也没有充分利用到问题的先验知识.

将问题的先验知识引入到 aiNet 算法中,并结合 RLAIS 和 EFAIN 算法,本文提出一种改进的模糊人工免疫网络数据分类方法(Fuzzy Artificial Immune Network Classification, FAINC),UCI 数据集的仿真实验结果表明,该方法不仅具有较高的分类正确率和数据浓缩率,而且更稳定、可靠.

1 分类问题描述

假设经过数据归一化后的样本集合为 $\mathbf{X}$,其中每个模式 $\mathbf{x}_i$ 有 p 个特征,它所对应的类别集合为 $\mathbf{S}$,如下所示

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{bmatrix}, \quad \text{其中 } m \leq n$$

在构造分类器之前要先随机地选取训练集合 $\mathbf{X}_{tr}$ 和测试集合 $\mathbf{X}_{te}$ ($\mathbf{X}_{tr} \subset \mathbf{X}$, $\mathbf{X}_{te} \subset \mathbf{X}$),一般地,集合 $\mathbf{X}_{tr}$ 和 $\mathbf{X}_{te}$ 中的样本包含所有的类别 $\mathbf{S}$.分类器的构造过程一般分为训练和测试两个阶段.在训练阶段,通过分析训练数据集 $\mathbf{X}_{tr}$ 的特点或者学习 $\mathbf{X}_{tr}$ 提供的信息,产生一个 $\mathbf{X}$ 与 $\mathbf{S}$ 的映射关系的准确描述或模型.在测试阶段,利用训练得到的描述或模型对测试数据集 $\mathbf{X}_{te}$ 进行分类,通过检验其分类正确率,判断

所构造的分类器是否有效.

如果能够通过训练得到 $\mathbf{X}$ 与 $\mathbf{S}$ 之间的准确函数关系,则可以解决该数据集的所有分类问题,然而这种准确函数关系往往很难找到.人工智能为相应分类器的构造提供了新思路^[2-4].

2 模糊人工免疫网络数据分类算法

2.1 算法原理

不论是 RLAIS 还是 aiNet 都存在一个共同的缺陷,那就是缺乏对先验知识的运用.模糊 C 均值(Fuzzy C-Means, FCM)聚类算法将样本在每个类的隶属度扩展到区间 $[0, 1]$,这和抗体、抗原识别的免疫机理很相似,所以本文用 FCM 获得数据的先验知识,并作为人工免疫网络的疫苗,通过克隆选择、网络压缩、免疫成熟、记忆等算子的处理形成各类对应的子网络,然后将这些子网络合并后得到一个完整的人工免疫网络,这个网络也就是样本数据的一个描述,最后根据“邻近原则”完成数据的分类任务.

依据文献[5, 11]算法的参数定义,本文令 $\mathbf{X}_{tr}$ 中每个状态 s_i 所对应的样本集合为 $\mathbf{X}_i$ ($\mathbf{X}_i \subset \mathbf{X}_{tr}$, $i=1, 2, \dots, m$),称为第 i 类; $\mathbf{M}_i$ 为第 i 个状态训练得到的记忆单元,即子网络; N 为初始种群规模; N_c 为最大克隆数; ζ 为选择成熟分子的比率; h 为被选择克隆的抗体数目; σ_s 、 σ_d 分别为网络压缩阈值和抗体-抗原亲和度阈值.为了更好地叙述算法,作如下定义和说明.

(1) 亲和度计算方法

$$f_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\|^2 \quad (1)$$

计算抗体-抗原亲和度时 $\mathbf{x}_i$ 表示抗原, $\mathbf{y}_j$ 表示抗体; 计算抗体-抗体亲和度时 $\mathbf{x}_i$ 和 $\mathbf{y}_j$ 都表示抗体.

(2) 网络压缩阈值 σ_s 的计算方法

$$\sigma_s = k_0 \bar{e} \quad (2)$$

式中: $\bar{e}$ 为状态 s_i 所对应的样本 $\mathbf{X}_i$ 的最小生成树中边的平均长度; k_0 为系数.

(3) 目标函数定义

首先需要指出的是,这里定义的目标函数只是为两次免疫应答抗体网络是否已经稳定提供一种停机判别,它不是最终分类的误差函数

$$J_i = \sum_{j=1}^{c_i} \sum_{l=1}^{n_i} \mu_j(\mathbf{x}_l^i) \|\mathbf{x}_l^i - \mathbf{a}_j^i\|^2 \quad (3)$$

$$i = 1, 2, \dots, m$$

式中: c_i 为第 i 类网络中抗体的数目; n_i 为 $\mathbf{X}_i$ 中的样本数目; $\mathbf{a}_j^i$ 表示第 i 类对应的子网络中的第 j 个

抗体; $\mu_j(x_l^i)$ 表示第*i*类中的第*l*个抗原对第*j*个抗体的隶属度。

(4)亲和度成熟算子

$$C_{\text{new}} = C_{\text{old}} - \alpha(C_{\text{old}} - x_l^i) \quad (4)$$

式中: C_{old} 是算法中经过克隆后得到的抗体集合; $\alpha = \exp(-1/f_l^i)$ 是学习率向量, f_l^i 是 C_{old} 与 x_l^i 的亲和度向量。

(5)抗体死亡算子

如果类别 s_i 所对应的免疫网络中一个抗体受到的抗原总刺激小于一个阈值 $N_{X_i-A_i}^i$,并且受到网络中其他抗体的总刺激小于阈值 $N_{A_i-A_i}^i$,则删除这个抗体,这就是抗体死亡算子

$$N_{X_i-A_i}^i = \frac{1}{c_i} \sum_{j=1}^{c_i} S_{X_i-a_j}^i \quad (5)$$

$$N_{A_i-A_i}^i = \frac{1}{c_i} \sum_{j=1}^{c_i} S_{A_i-a_j}^i \quad (6)$$

$$S_{X_i-a_j}^i = \sum_{l=1}^{n_i} \exp(-\|x_l^i - a_j^i\|) \quad (7)$$

$$S_{A_i-a_j}^i = \sum_{q=1, q \neq j}^{c_i} \exp(-\|a_q^i - a_j^i\|) \quad (8)$$

2.2 算法流程

步骤1:设定参数 $k_0, N, \sigma_d, \zeta, N_c, h$;

步骤2:对于每一个 $X_i, i=1, 2, \dots, m$

(1) $k=0$,按照式(2)计算网络压缩阈值 σ_s ,用FCM算法预处理数据,得到的 N 个聚类中心作为疫苗 B ,并将 B 作为初始抗体群 $A_i(k)$;

(2)对于每一个抗原 $x_l^i, x_l^i \in X_i$,完成以下操作

① 用式(1)计算 x_l^i 与 $A_i(k)$ 的亲和度;

② 选择亲和度最低的 h 个抗体,按照亲和度越低,抗体规模越大的原则进行克隆,最大克隆数为 N_c ,得到抗体集合 C_{old} ;

③ 按照式(4)执行抗体成熟算子得到 C_{new} ;

④ 计算 C_{new} 和抗原 x_l^i 的亲和度,并将亲和度最低的 ζ 个抗体存入 R 中;

⑤ 令 R 中与 x_l^i 亲和度大于门限 σ_d 的记忆抗体死亡;

⑥ 用式(1)计算 R 中抗体-抗体间亲和度,令亲和度小于 σ_s 的记忆抗体死亡;

⑦ $M_i \leftarrow [M_i; R]$;

(3)分别按照式(7)和(8)计算 M_i 中记忆抗体的刺激值,按照式(5)和(6)计算阈值,执行抗体死亡算子;

(4)更新抗体群, $k=k+1, A_i(k) \leftarrow [M_i; B]$;

(5)按式(3)计算误差目标函数;

(6)如果满足停止条件则继续,否则转步骤2(2);

(7)以 M_i 为初始聚类中心,用FCM聚类算法来调整 M_i ;

(8)对 M_i 执行抗体死亡算子,最终得到状态 s_i 所对应的子网络;

步骤3:链接各子网络得到最终网络 M ;

步骤4:根据“邻近原则”完成分类。

步骤2(6)中的停止条件由最大迭代代数和误差目标函数共同确定,当两代误差目标函数之差小于某一值或者满足最大迭代代数时停机。

3 实验与讨论

3.1 实验结果

选用 Iris、Wine、Ionosphere、Diabetes 等 4 个 UCI 数据集做测试。本文对每个数据集都进行 50 次独立实验,在每次实验中,随机选择每类数据中的一半作为训练数据,其余作为测试数据。实验结果如表 1 所示,表中的分类正确率 α 和浓缩率 β 分别采用式(9)和(10)计算

$$\alpha = \frac{n_r}{n_{te}} \times 100\% \quad (9)$$

式中: n_r 为正确分类的数据数目; n_{te} 为测试数据总数目。

$$\beta = \left(1 - \frac{n_A}{n_{tr}}\right) \times 100\% \quad (10)$$

式中: n_A 为最终网络中的抗体数目; n_{tr} 为训练数据总数目。

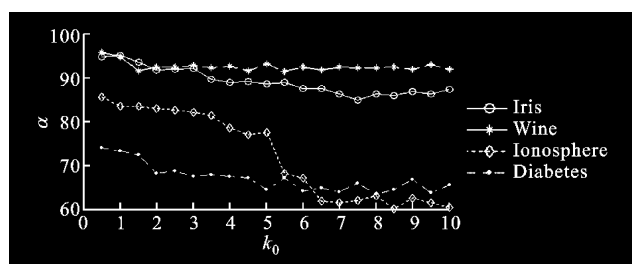
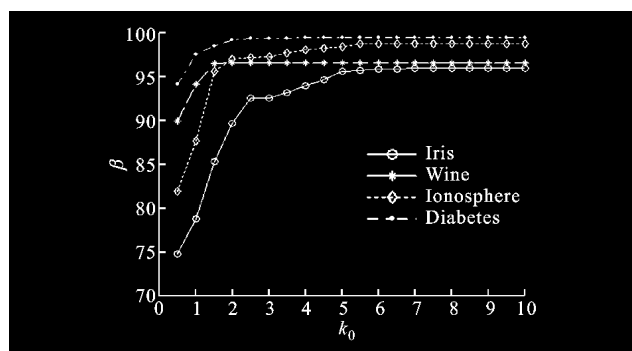
表 1 中的 FAINC 为本文中的分类算法,算法中的一些参数设定为 $k_0=1.5, \sigma_d=0.5, N=10, h=4, N_c=10, \zeta=0.2$; aiNet 栏的结果为以 aiNet 算法为网络训练算法,然后用“邻近原则”得到的分类结果。为了增加可比性,采用与 FAINC 算法相同的方法计算 aiNet 中 σ_s 的值。aiNet 中其他各参数分别为 $k_0=1.5, \sigma_d=1.0, N=10, h=4, N_c=10, \zeta=0.2$,即除了 σ_d 外,其他各参数的取值与 FAINC 算法相同。从分类正确率来看,FAINC 在 4 组数据集上的平均分类正确率比 aiNet 平均高出 7.26%,而方差却小 1.4%,在数据浓缩率上,FAINC 平均比 aiNet 高出 11.16%,而方差却比 aiNet 小 1.88%。这说明与算法机理相似的 aiNet 相比,FAINC 不仅有更好的分类效果,而且稳定性更高、更可靠。

表1 实验结果

方法	统计量	Iris	Wine	Ionosphere	Diabetes
FAINC	α	平均值	96.08	93.91	84.19
		方差	1.56	2.76	2.71
	β	平均值	85.81	96.14	96.93
		方差	1.92	0.65	0.55
aiNet	α	平均值	88.03	93.18	68.89
		方差	6.07	2.80	3.29
	β	平均值	71.15	89.00	83.94
		方差	5.58	1.97	2.24

3.2 参数影响情况

大量实验表明,在FAINC算法中,参数 N 、 σ_d 、 ζ 、 N_c 和 h 对算法性能影响不大,因此本文只讨论 k_0 取值的影响.取 $\sigma_d=0.5$, $N=10$, $h=4$, $N_c=10$, $\zeta=0.2$,当 k_0 从0.5到10以步长0.5变化时,其对3.1小节中测试问题准确率和浓缩率的影响如图1所示.

(a) k_0 对分类正确率的影响(b) k_0 对分类浓缩率的影响图1 k_0 对算法性能的影响情况

上述结果表明, k_0 对正确率和浓缩率都有较大的影响,4个数据集的分类正确率都随着 k_0 的增大而逐渐减小,而浓缩率却随着 k_0 的增大而逐渐增大并趋于一个固定值.这是因为 k_0 通过影响网络压缩

阈值 σ_s 作用于最终形成的网络结构, k_0 越大网络压缩越严重,最终网络中抗体数越少,使网络不能充分概括数据信息,导致正确率下降.从图1b中可以看到,在本实验中只有当 $k_0 < 3$ 时,才对浓缩率有比较大的影响,所以在保证浓缩率的情况下,应取相对小的 k_0 值.

4 分析与结论

由于FAINC算法中抗体死亡算子的作用,使噪声对结果的影响比aiNet更低,另外在FAINC中还有用FCM算法微调抗体位置的步骤,这让最终网络中的抗体与主数据流的形状更加吻合,所以分类效果要比aiNet好.依据文献[12]的分析,FAINC算法的时间复杂度约为 $O(N^2)$,其中 N 为训练样本数.通过实验可以得到如下结论:

(1)从数据本身出发,通过浓缩来完成分类的方法能够有效地解决数据分类问题;

(2)人工免疫网络具有很高效的数据浓缩能力,FAINC首先利用人工免疫网络对数据进行浓缩,然后再用FCM调整抗体位置,使得到的抗体更能代表样本数据;

(3)FAINC不论在正确率还是数据浓缩率上都要高于aiNet,而且更加稳定、可靠.

本文方法还存在一些不足:首先是基于欧式距离的亲亲和度的计算方法限制了算法对高维数据的处理能力;其次,式(2)中系数 k_0 决定着 σ_s ,继而决定着最终的网络结构,如果在整个训练过程中将 k_0 取为一个定值则忽略了各类样本的差异性,会影响到最终的正确率和数据浓缩率.所以,怎样构造更合理的亲和度计算方法和怎样根据抗原的分布情况自适应地选择系数 k_0 ,将是一个值得探讨的问题.

参考文献:

- [1] 刘红岩,陈剑,陈国青.数据挖掘中数据分类算法综述[J].清华大学学报(自然科学版),2002,42(6):727-730.
Liu Hongyan, Chen Jian, Chen Guoqing. Review of classification algorithms for data mining [J]. Journal of Tsinghua University (Science and Technology), 2002,42(6):727-730.
- [2] Chiu Chaohang, Hsu P L. A constraint-based genetic algorithm approach for mining classification rules[J]. IEEE Transactions on Systems, Man and Cybernetics, 2005,32(2):205-220.
- [3] Rustkowski L. Adaptive probabilistic neural networks

- for pattern classification in time-varying environment [J]. *Neural Networks*, 2004, 15(4): 811-827.
- [4] Secker A, Freitas A A, Timmis J. AISEC: an artificial immune system for e-mail classification [C]// *Proceedings of 2003 Congress on Evolutionary Computation*. Piscataway, USA: IEEE Press, 2003: 131-138.
- [5] 杜海峰,王孙安. 基于 ART-人工免疫网络的数据浓缩方法研究[J]. *模式识别与人工智能*, 2001, 14(4): 401-405.
- Du Haifeng, Wang Sun'an. Data enriching based on ART-artificial immune network [J]. *Pattern Recognition and Artificial Intelligence*, 2001, 14(4): 401-405.
- [6] Fu Dongmei, Zheng Deling. Designing and analysis of an immune controller based on the improvement VARELA immune network model [C]// *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*. Piscataway, USA: IEEE Press, 2005: 1121-1126.
- [7] Castro L N, Zuben F J V. An evolutionary immune network for data clustering [C]// *Proceedings of the Sixth Brazilian Symposium on Neural Networks*. Piscataway, USA: IEEE Press, 2000: 84-89.
- [8] D'haeseleer P, Forrest S, Helman P. An immunological approach to change detection: algorithms, analysis and applications [C]// *Proceedings of the 1996 IEEE Symposium on Research in Security and Privacy*. Piscataway, USA: IEEE Press, 1996: 110-119.
- [9] 焦李成,杜海峰. 人工免疫系统进展与展望[J]. *电子学报*, 2003, 31(10): 1540-1548.
- Jiao Licheng, Du Haifeng. Development and prospect of the artificial Immune system [J]. *Acta Electronica Sinica*, 2003, 31(10): 1540-1548.
- [10] Timmis J, Neal M. A resource limited artificial immune system for data analysis [J]. *Knowledge Based Systems*, 2001, 14(3/4): 121-130.
- [11] 谢明,杜海峰,焦李成. 一种用于数据聚类的进化模糊人工免疫网络[J]. *计算机科学*, 2004, 31(10A): 219-223.
- Xie Ming, Du Haifeng, Jiao Licheng. A novel evolutionary fuzzy artificial immune network for data clustering [J]. *Computer Science*, 2004, 31(10A): 219-223.
- [12] 焦李成,杜海峰,刘芳,等. 免疫优化计算、学习与识别 [M]. 北京:科学出版社, 2006.

(编辑 荆树蓉)